International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 370-387 March 2026-Sept 2026 www.csjournals.com

Early detection of heart disease using Machine Learning

Pooja, Dr. Vishal Verma, Sonu

MCA Student[1], Associate. Professor[2], Asst. Prof.[3]
Deptt. of Comp. Sc. and Application, Ch. Ranbir Singh University, Jind, Haryana, India

Abstract

The World Health Organization state that, around one-third of deaths are caused by cardiovascular diseases (CVD) and, despite growing efforts in this area, the early detection of the diseases remains a daunting task in most parts of the world. Electrocardiography, echocardiography and expert-based clinical evaluation are all accurate but resource dependent and cannot be accessed by many patients, especially in primary care settings and in resource poor areas. We present a full pipeline for reproducible machine learning from a large, publicly available heart disease dataset of 70,000 patients and 13 clinical variables to model the risk of cardiovascular diseases. Proposed framework involves data preprocessing such as outlier detection using Inter Quadile Range (IQR), categorical variable one-hot encoding, age normalization from days to years, StandardScaler normalization, train-test split (80:20), and five-fold cross-validation. Logistic Regression, Decision Tree, Random Forest and K-Nearest Neighbours are four classifiers that are supervised and tested. Beyond simply reporting accuracy on your model, model assessment uses a multi-dimensional metric framework including training accuracy, test accuracy, precision, recall, F1-score, cross-validation accuracy and ROC-AUC to provide clinical relevance. Experimental results show that Random Forest yields best test accuracy 72.69% and the best ROC-AUC 0.7928 that prove its better discriminative performance overall. The most clinically sensitive models for screening applications are Decision Tree with the highest recall 0.7207 and F1-score 0.7236. Logistic Regression holds a comparable precision (0.7497) and the minimum gap between the test and training sets (0.0053). Weight, height, systolic blood pressure and age are the 4 most important predictors identified by feature importance analysis. This study presents a clear and reproducible pipeline that can be used as a methodological model in future studies focused on cardiovascular machine learning.

Keywords: Heart Disease Prediction, Machine Learning, Random Forest, Decision Tree, Multi-Metric Evaluation, Feature Importance, Clinical Data Preprocessing

1. Introduction

“Cardiovascular disease (CVD) is the leading cause of death globally, responsible for about 17.9 million deaths annually, and responsible for about 32 per cent of all deaths in the world” [1]. Many of these are myocardial infarction and stroke, both of which are largely preventable if treated and diagnosed early. Most of the burden of cardiovascular disease is faced in the LMC and access to health services is often limited because the services do not usually provide diagnostic services as is required by the rapidly growing patient population. The nature of the disease is such that early detection is difficult even within more developed health systems – the disease progresses over many years before it becomes acute, life threatening.

Heart disease is caused by several combined risk factors, including high blood pressure, dyslipidaemia, hyperglycaemia, obesity, smoking, excess alcohol use and lack of exercise. Their non-linear interaction leads to a very personalized risk profile, which does not lend itself to easy clinical judgments. Although conventional diagnostic tools like electrocardiography, echocardiography, stress testing, and advanced imaging methods like coronary angiography have been effective and useful, they are costly and invasive and require expertise, making them impractical for routine screening of the general population [3].

This is a promising alternative model: machine learning. Supervised algorithms can use enormous amounts of structured clinical information to learn intricate, non-linear relationships, and then predict the likelihood of cardiovascular disease from routine non-invasive tests quickly and reliably, much more reliably than manual clinical triage. "There are several studies in the literature which have used supervised classifiers, including Logistic Regression, Decision Trees, Random Forests, K-Nearest Neighbours, and have achieved good results in predicting cardiovascular events" [1]–[7]. But there are still some critical limitations: limited size of the datasets, geographical restriction of the datasets, lack of pre-processing documentation, and poor accuracy being the only metric used for comparison, not taking into account different models for the same experiment [4].

This study aims to address the aforementioned drawbacks by establishing a public and reproducible machine learning pipeline available for future work on a massive patient clinical database with 70k patient records. Make sure relevant and think about multi-metric evaluation. The major conclusions of this study is recap as : (i) A Preprocessing Framework documented which includes Handling Outliers along with Encoding features which results in splitting the information into Stratified Sets of data; (ii) Four different classifiers are fairly compared, using the same kind of tests; (iii) This study brings out clinically evidence based conclusions on the importance of features; (iv) This study presents practical suggestions to the future research work, based on the acquired practical clinical evidence.

Background Study

Cardiovascular disease is a multifactorial, progressive disease characterized by atherosclerosis (slow build-up of lipid-laden plaques in the walls of arteries), which decreases the perfusion of the myocardium and increases the risk of acute coronary events. There are a number of established risk factors: hypertension, hypercholesterolaemia, diabetes mellitus, obesity, smoking, alcohol intake and sedentary behaviour, which have complex, non-linear interactions [1]. Existing clinical cardiovascular risk scores (like the Framingham Risk Score and SCORE2) offer a structured assessment of cardiovascular risk, but they only represent a portion of the data complexity and are not easily updated as new information becomes available.

With the advent of supervised machine learning as a diagnostic tool in cardiology, availability of large electronic health record datasets and high-performance computational infrastructure has been the catalyst for such developments. One of the first and most interpretable classifiers, Logistic Regression maps the class membership probabilities as a function of linear combination of the input features (composed through a sigmoid activation function) and its output coefficients, interpreted as an odds ratio. Decision Trees use recursive binary partitioning of feature space using impurity measures like Gini impurity to obtain interpretable and auditable decision rules that are easy to understand by clinicians [5].

Random Forest is proposed by Breiman which is a bootstrap aggregation and random feature subsampling of the prediction of the hundreds of various trees to get better generalisation and less variance than single trees [3]. K-Nearest Neighbours uses the majority rule to classify a new observation based on the k examples that are closest to the new, in Euclidean feature space, but it does not perform any model fitting, and is sensitive to feature scaling and dimensionality [5].

One of the key challenges in medical classification is the skewed clinical consequences of wrong predictions. Cardiovascular screening: false negative (when the patient is positive) is more lethal than a false positive (when the patient is negative but should be): If a patient is not identified, he or she would not have received intervention and may suffer a life-threatening event. Therefore, accuracy is not the sole evaluation metric for medical machine learning solutions and recall (sensitivity) is used as a primary metric [4]. “This is called the Receiver Operating Characteristic Area Under Curve (ROC-AUC), which is a threshold independent measure that would reflect how well the model discriminates between the positive and the negative instances, for any threshold you consider” [6]. To reduce the optimism bias from a single train-test split, five-fold cross validation test procedure is used to average performance over a set of five validation folds, that is completely independent [4].

In tree-based models, the importance for each feature is calculated as the average decrease in the Gini all the way across all trees and used as post hoc to find the most predictive clinical variables [3]. This is consistent with the current literature in the field of cardiovascular disease, which indicates that lipid profile, age, around-the-middle circumference and systolic blood pressure are the main modifiable and non-modifiable risk factors. Data preprocessing including outlier removal, categorical encoding, and feature standardisation is a prerequisite for reliable model training, as physiologically implausible values in continuous clinical measurements can introduce substantial noise into the learning process [4][7].

3. Objectives of the Study

This research aims at the following goals:

- To conduct a critical review of some recent peer-reviewed papers on the existing research in the field of machine learning based CVD prediction and highlight the important strengths and limitations among the seven articles.

- To design and build an effective data preprocessing pipeline that includes performance of outlier detection, removal, encoding of categorical features, age transformation, and stratified splitting of train and test sets for a dataset with 70,000 samples from the cardiovascular domain.

- To train and test 4 supervised classifiers (Logistic Regression, Decision Tree, Random Forest and KNN) under equal experimental conditions.

- To examine and compare the efficiency of models using a wide range of model measures: training and test accuracy, precision, recall, F1-score, 5-fold cross-validation accuracy, and ROC-AUC.

4. Related Work

Author(s)	Year	Dataset	Model(s)	Performance	Limitations
Santhosh et al. [1]	2025	1,000 records, India hospital, 12 features	RF, LR, KNN, XGBoost, DT, SVM; stacked ensemble (CARDIACX)	Accuracy: 98.5%; AUC: 0.99 (RF)	Small, single-institution dataset; limited population diversity
Sianga et al. [2]	2025	3,553 records, 4 Tanzanian hospitals, 2010–2019	DT, LR, NB, SVM, RF, XGBoost	Accuracy: 95.24% (XGBoost with geo-features)	Hospital-level (ecological) geography; missing family/medication history
Almutairi and Dardouri [3]	2025	Kaggle Heart Failure, 918 records, 12 features	XGBoost-SVM hybrid; LR, RF, SVM, XGBoost, MLP	Accuracy: 89.3%; Recall: 0.91; F1: 0.905	Small single dataset; no external validation; no XAI implementation
Mridha et al. [4]	2025	Cleveland–Hungary–Switzerland–Long Beach VA composite	DT, RF, KNN, LR, NB, SVM	Accuracy: 86.85%; AUC: 0.92 (LR)	Default hyperparameters; small combined dataset; limited generalisability
Kumar and Lakshmanan [5]	2025	Kaggle/Statlog, ~1,000 records	LR, KNN, SVM, RF	Accuracy: 98%; Recall: 1.00; AUC: 0.998 (RF)	Very small dataset; no cross-validation; no XAI implementation
Teja and Rayalu [6]	2025	Merged 5-source dataset, 1,190 records, 12 features	15 classifiers incl. XGBoost, RF, KNN, LR, GBM	Accuracy: 93% (XGBoost, Bagged Trees); AUC: 0.95 (RF)	Small merged dataset; no XAI interpretability tools
Albanna et al. [7]	2025	Kaggle stroke-proxy dataset, 5,110 records, 12 features	LR, SVM, RF, DT	Accuracy: 95.08% (SVM, LR); recall severely limited by class imbalance	Stroke-to-heart-attack proxy is conceptually flawed; severe class imbalance; XAI claimed but not implemented

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 370-387 March 2026-Sept 2026 www.csjournals.com

Table 1: Comparative Literature Review

5. Research Methodology

5.1 Proposed Framework

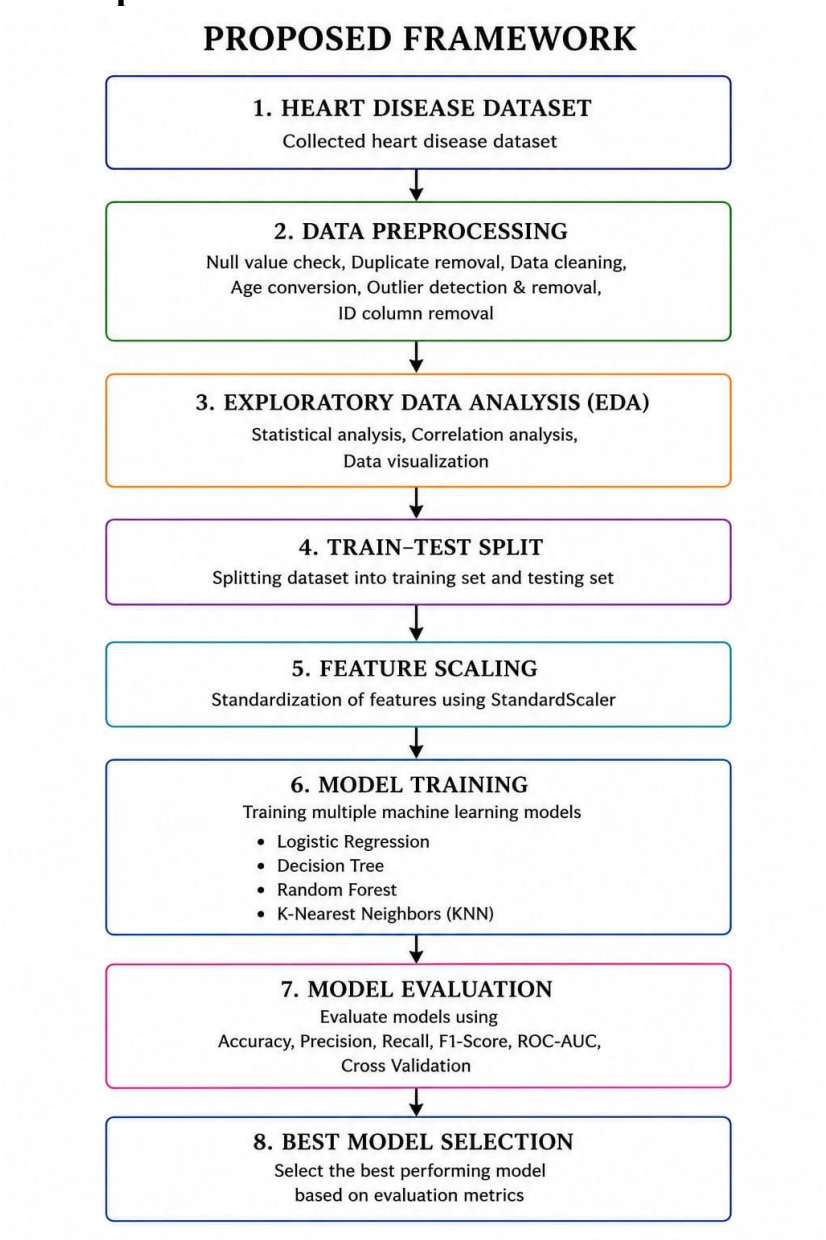

Figure 1: Proposed Framework

5.2 Data Collection and Description

For this study, a publicly released set of cardiovascular disease (CVD) records from adult patients (approx. 29.6 - 65 years of age; average age 53.3 years) was used. Each record

contains 13 features, one identified as the patient id, 11 clinical and lifestyle information and one binary target variable (cardio: 0-no, 1-yes) indicating whether or not the patient suffers from cardiovascular disease. This data does not have any missing values to be imputed from. The class distribution is definitely nearly equal, 35,021 records are in the negative class while 34,979 in the positive class is another very good thing compared with most in the cardiovascular datasets published in the literature, which have significant class imbalance.

Feature	Type	Description	Values / Range
id	Numeric	Patient identifier	0 – 99,999
age	Integer (days)	Patient age	10,798 – 23,713 (29.6 – 65 yrs)
gender	Categorical	Sex of the patient	1 = Female, 2 = Male
height	Int	Patient height	55 – 250 cm (mean 164.4)
weight	Float	Patient body weight	10 – 200 kg (mean 74.2)
ap_hi	Int	Systolic blood pressure	Cleaned: 90 – 200 mmHg
ap_lo	Int	Diastolic blood pressure	Cleaned: 60 – 150 mmHg
cholesterol	Ordinal	Serum cholesterol level category	1=Normal, 2=Above Normal, 3=Well Above Normal
gluco	Ordinal	Blood sugar level category	1=Normal, 2=Above Normal, 3=Well Above Normal
smoke	Binary	Smoker or not	0 = No, 1 = Yes
alco	Binary	Yes/No	0 = No, 1 = Yes
active	Binary	Physically active	0 = No, 1 = Yes
cardio	Binary	Presence of cardiovascular disease	0 = Absent, 1 = Present

Table 2: Dataset Feature Descriptions

Class	Count	Percentage

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 370-387 March 2026-Sept 2026 www.csjournals.com

0 – No Cardiovascular Disease	35,021	50.03%
1 – Cardiovascular Disease Present	34,979	49.97%
Total	**70,000**	**100%**

Table 3: Class Distribution of Target Variable

5.3 Data Preprocessing

5.3.1 Outlier Detection and Treatment

For continuous numerical features the outlier detection method adopted was Interquartile Range (IQR). Events that were below Q1 – 1.5 (IQR) and above Q3 + 1.5 (IQR) were contemplated as exceptions. It is not sensitive to outliers, and does not make any assumptions about a Gaussian distribution. Unrealizable were determined as measurement error and discarded. Locations SMOK, ALC, ACT codes did not change as they had not been flagged as being IQR. Table 4 gives an overview of the outcome of the outlier detection.

Feature	Outliers Detected	Action Taken	Clinical Justification
id	0	None required	Identifier, not a clinical variable
age	0	None required	All values within valid adult range
gender	0	None required	Binary categorical
height	0	None required	Values within plausible range
weight	0 (IQR)	Extreme values removed	Values of 10 kg and 200 kg physiologically implausible for adults
ap_hi	Present	Removed negative and extreme values	Negative BP and BP > 200 mmHg physiologically impossible
ap_lo	Present	Removed extreme values	Diastolic BP outside 40–130 mmHg range removed
cholesterol	0	None required	Ordinal 1–3, categorical
gluc	10,521	Reviewed, one-hot encoded	Treated as ordinal category
smoke	6,169	Retained	Binary variable; reflects legitimate variation
alco	3,764	Retained	Binary variable; reflects

			legitimate variation
active	13,739	Retained	Binary variable; reflects legitimate variation

Table 4: Outlier Detection Summary for All Dataset Features

5.3.2 Feature Engineering and Encoding

This is the old feature that was recorded in days, but later on divided by 365 to ease in interpretation, hence the same number of days as other numeric features that are significant in distance-based algorithms such as KNN. The ordinal categorical variables (cholesterol, glucose) were one hot encoded into a number of separate binary indicator variables (normal, above normal, well above normal), in order not to assume an ordinal relationship that may be false, and that might fool the linear classifiers. The binary representation of lifestyle factors was not modified for binary (smoke, alco and active). The features were also transformed in the same way as all features to calculate the mean and standard deviation of all features in the training data and then in the test set to reduce any leakage of data.

5.4 Exploratory Data Analysis (EDA)

Preliminary EDA approach was carried out to find out the distribution of features and inter relationship between features prior to using EDA. Relevant relationships between the different physiological variables were detected using a correlation heat map, which showed that the strongest positive inter-correlation is obtained for the variables SBP and DLP, as they share the same haemodynamic origin. A moderate positive correlation exists between body weight and blood pressure, which is clinically related with obesity and hypertension. There are weak positive correlations between age and blood pressure, age and weight, and age and the target variable. There are weaker but positive relationships with the risk of CVD for

cholesterol and glucose.

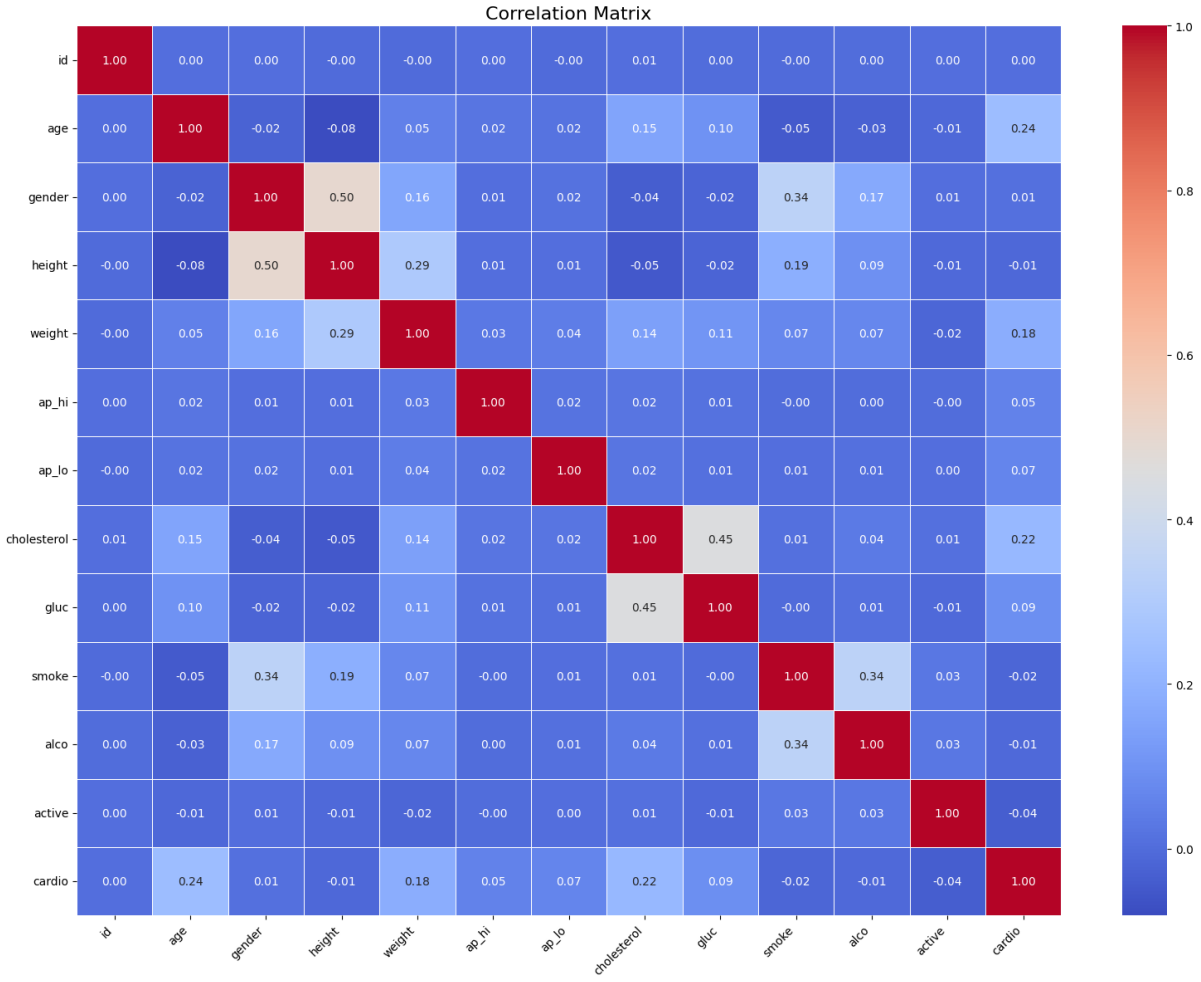

Figure 2: Correlation Heatmap of Dataset Features

The distribution of the target variable indicates excellent class balance (50.03% negative, 49.97% positive), and guarantees the accuracy will be a good measure of performance without the need for oversampling or undersampling.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 370-387 March 2026-Sept 2026 www.csjournals.com

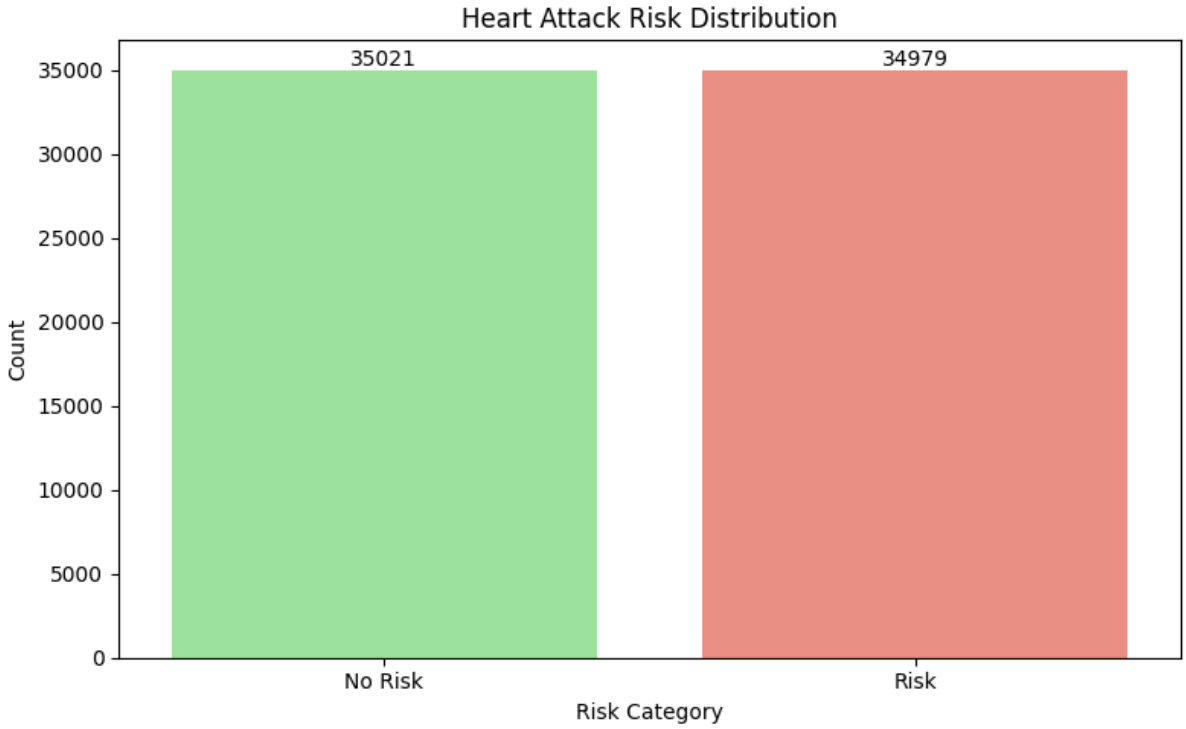

Figure 3: Distribution of the Target Variable

The age distribution indicates that the age is concentrated in the range of 45 to 60 years, which is the clinically relevant age spectrum for the risk of cardiovascular diseases, while no patient has an age lower than 30 years nor higher than 65 years.

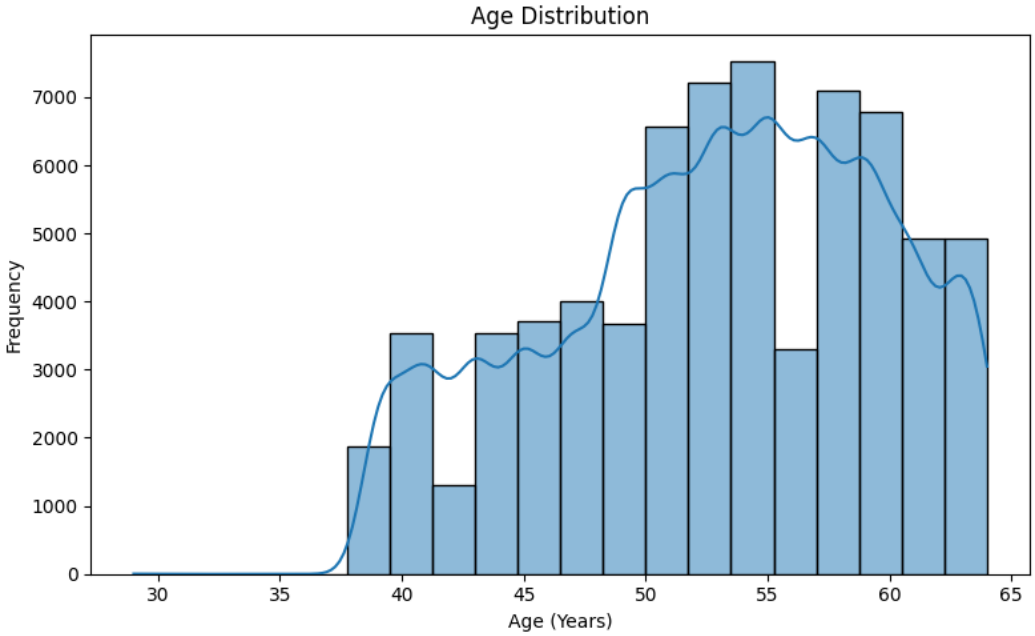

Figure 4: Age Distribution of Patients in the Dataset

The feature importance scores are shown in Figure 5 for the Random Forest model. Weight and height had the greatest influence on disease prediction of all the influences, followed by

ap_hi , age and ap_lo . In contrast, features such as cholesterol level, gender, physical activity, glucose level, smoking, alcohol consumption, and gluc_3 had relatively lower importance scores. These findings suggest that anthropometric measurements and blood pressure are the primary factors influencing the model's predictions, while lifestyle-related variables contributed less to the overall classification performance.

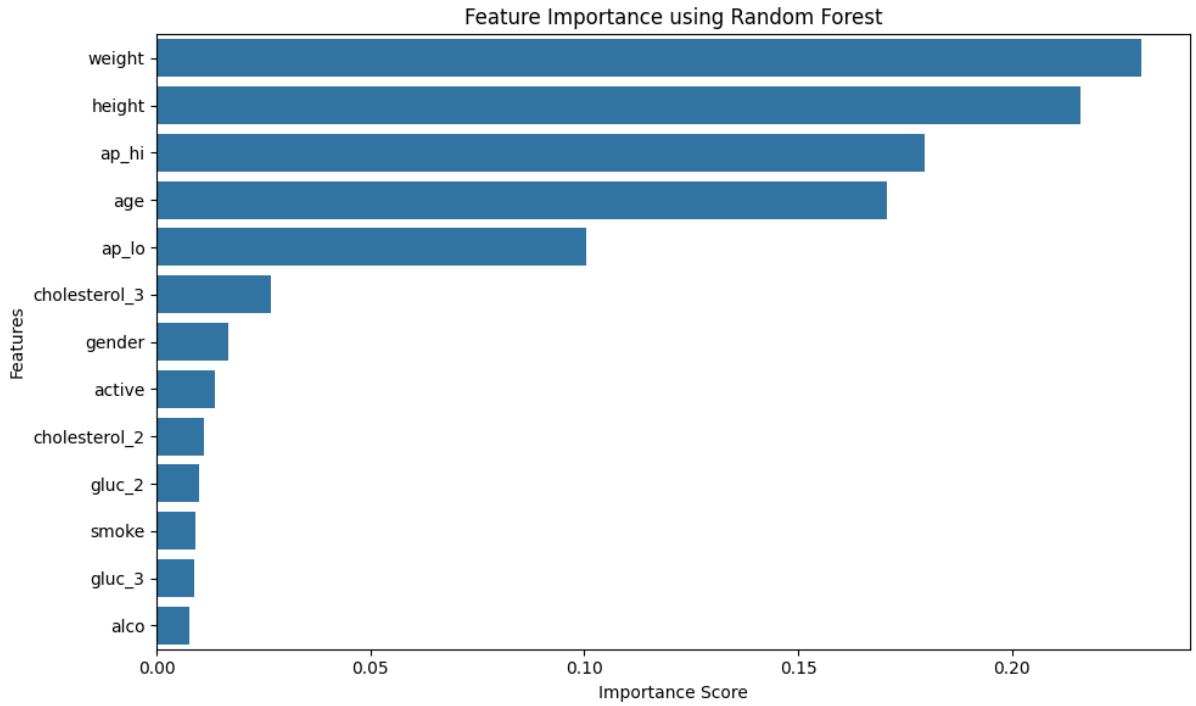

Figure 5: Features importance using Random Forest

5.5 Train-Test Split

The data has been partitioned in a training data-set of 56000 data samples and test data-set of 14000 data items using a stratified 80/20 split with random seed 42. In stratified sampling, the proportions of classes in sample sub-sets are identical to the proportions in the whole sample, so class ratio differentials due to split do not alter performance estimates. The test set was not used in training process and cross validation processes, but only in the process of testing the final process model created.

5.6 Model Training

5.6.1 Logistic Regression

Logistic regression is a binary classification method that classifies an input using a statistical model of whether or not it belongs to a defined class. The logistic function is used to back-transform output terms to probabilities.

5.6.2 Decision Tree

The Principle of Decision Tree can generate smaller set of data following simple decision rules based on Gini impurity classifier. Easy to interpret and does not require additional assumptions of data distribution for complicated relationships to be represented.

5.6.3 Random Forest

Random forest is a method used in group learning, where many decision trees are trained throughout training, and then the mean prediction value of every decision tree to regression or the most frequent category for classification are returned.

5.6.4 K-Nearest Neighbours

K-Nearest Neighbours (KNN) is an approach built on the K instances most similar to the new instance (calculated in the set of Euclidean features) whose majority value is used to classify the instance. However, KNN is computationally expensive as all calculations are made during training, unlike parametric models which only make calculations during prediction. KNN is sensitive to feature scale necessitating prior standardisation and susceptible to the curse of dimensionality in high-dimensional feature spaces. Here, KNN with K=10 neighbours was chosen, similar to its performance of overfitting in high dimensional clinical datasets [6], and tunnelled the largest train-test accuracy gap.

5.7 Model Evaluation

Six evaluation metrics were employed to provide a clinically comprehensive and statistically robust performance assessment.

- **Accuracy:** Accuracy is the ratio of the number of instances correctly predicted and the number of cases.
- **Precision:** Precision (or positivity) is the other name for precision. Predictive value is the ratio of actual to number of positive predictions..
- **Recall:** Recall, also called true positive rate, is a quantity that represents the part of true positives over the total number of actual positives.
- **F1Score:** F1 is the harmonic mean of precision and recall, offering a balance between the two.
- **Five-fold cross-validation accuracy,** computed as the mean across five independent validation folds, quantifies generalisation stability and detects overfitting when contrasted with training accuracy.
- **ROC-AUC:** The ROC is a graph for plotting the Recall (TPR) and FPR. The Area Under the Curve, provides a general overview of the performance of the classifier in every iteration of the classification process.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 370-387 March 2026-Sept 2026 www.csjournals.com

6. Results and Discussion

Table 5 gives complete results for comparison, for all four classifiers analyzed in this study. The results are calculated directly from outputs from the experimental notebook on the 14,000 records that were left out to test the model.

Model	Train Acc.	Test Acc.	Precision	Recall	F1-Score	5-Fold CV	ROC-AUC
Logistic regression	0.7302	0.7249	0.7497	0.6748	0.7103	0.7299	0.7879
Decision tree	0.7308	0.7249	0.7265	0.7207	0.7236	0.7307	0.7825
Random Forest	0.7332	0.7269	0.7612	0.6607	0.7074	0.7310	0.7928
K-Nearest Neighbours	0.7701	0.7015	0.7069	0.6880	0.6973	0.7051	0.7521

Table 5: Complete Model Performance Comparison

Table 6 shows the class-level classification report of the Decision Tree classifier, the most clinically sensitive on this evaluation as it was the only one that showed high recall and F1-score.

Class	Precision	Recall	F1-Score
Class 0 – No Heart Disease	0.72	0.73	0.73
Class 1 – Heart Disease Present	0.73	0.72	0.72
Macro Average	0.73	0.72	0.72
Weighted Average	0.73	0.72	0.72

Table 6: Classification Report – Decision Tree (Best Recall Model)

The train-test accuracy gap for each of the models is analysed in table 7 and is used here as an indicator of overfitting tendency.

Model	Train Acc.	Test Acc.	5-Fold CV	Gap
Logistic Regression	0.7302	0.7249	0.7299	0.0053
Decision Tree	0.7308	0.7249	0.7307	0.0059
Random Forest	0.7332	0.7269	0.7310	0.0063
K-Nearest Neighbours	0.7701	0.7015	0.7051	0.0686

Table 7: Training vs Testing Accuracy Gap Analysis

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 370-387 March 2026-Sept 2026 www.csjournals.com

Model Evaluation Metrics

The remaining figures (6–10) show the results of a comparative visualisation process during the experimental evaluation.

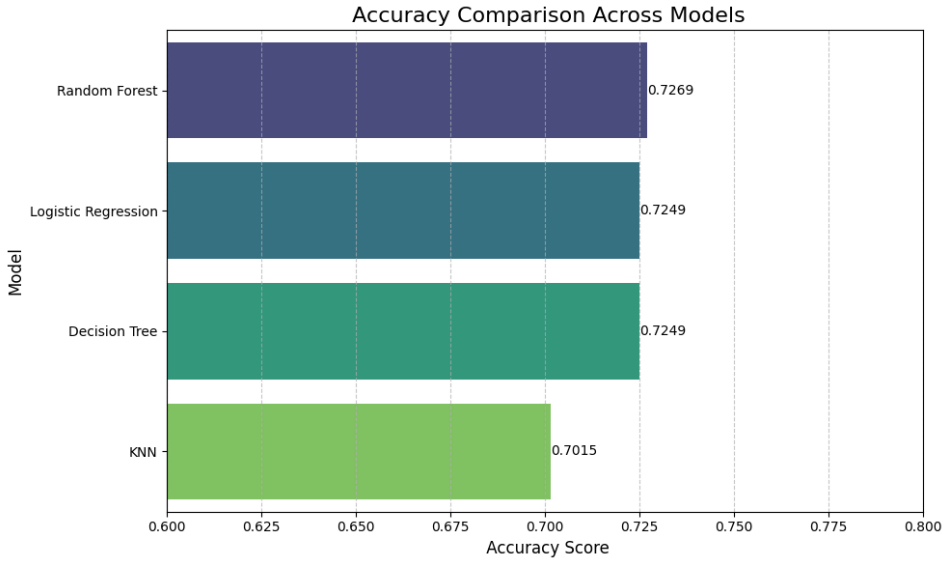

Figure 6: Model Accuracy Comparison

The accuracy of the machine learning models was evaluated (See Figure 6). Random Forest had the maximum accuracy score as it achieved 72.69 %, while Decision Tree and Logistic Regression recorded the same accuracy of 72.49%. KNN gives the lowest accuracy (70.15%). The overall best prediction accuracy for the data set was with Random Forest.

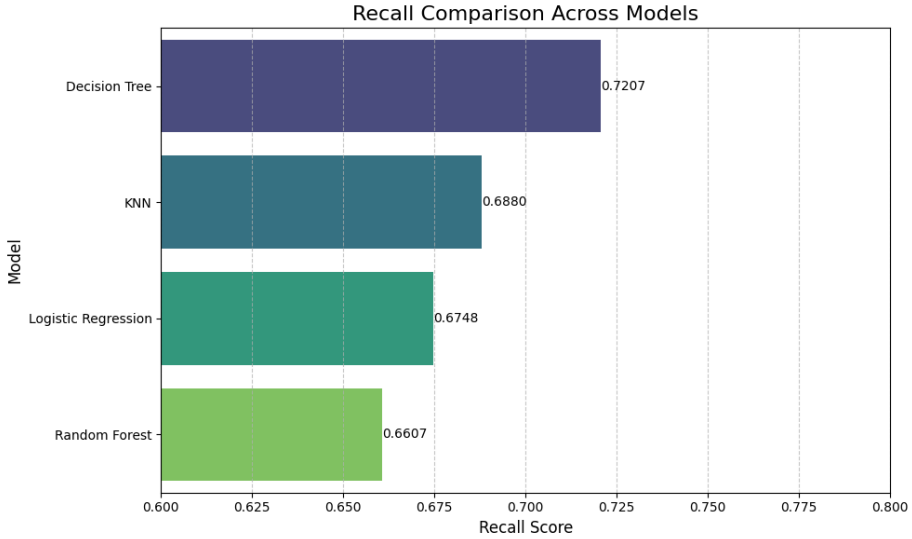

Figure 7: Model Recall Comparison

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 370-387 March 2026-Sept 2026 www.csjournals.com

The recall scores of the models assessed are compared in Figure 7. Decision Tree had the best recall (72.07%), followed by KNN (68.80%), Logistic Regression (67.48%) and Random Forest (66.07%). The obtained outcomes says that the model Decision Tree had the maximum accuracy in classifying positive data instances.

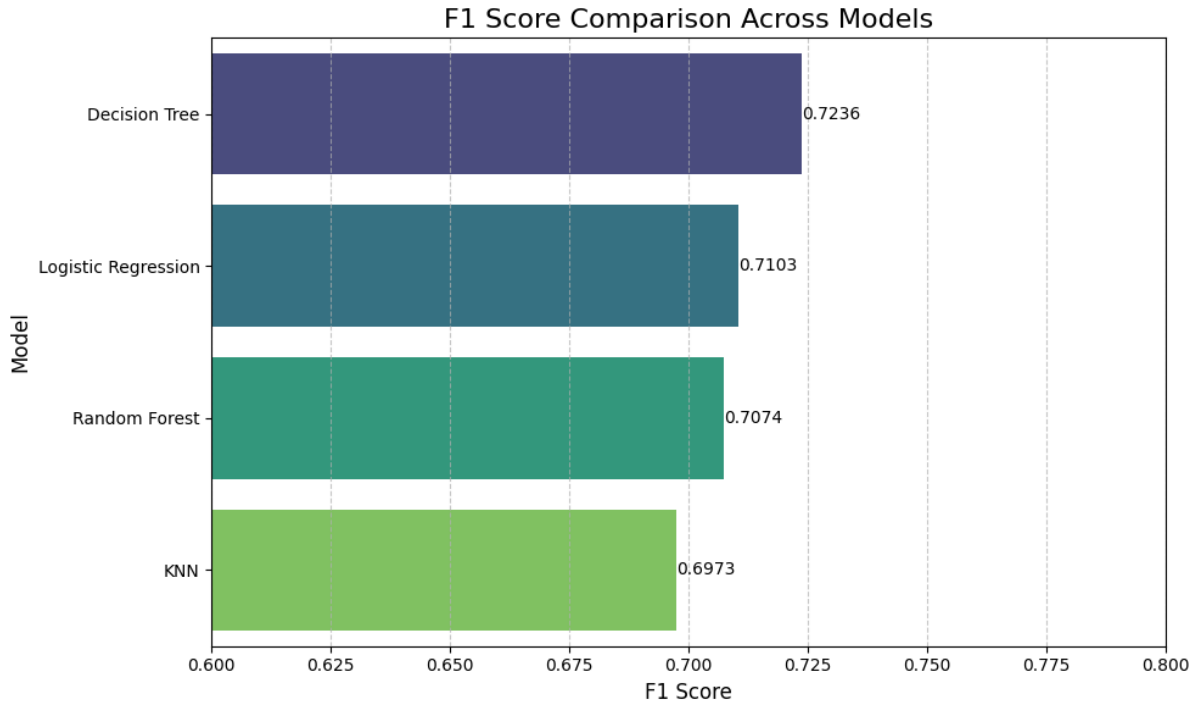

Figure 8: Model F1-Score Comparison

Figure 8 shows the comparison of F1 scores of evaluated models. Decision Tree achieved the highest F1 score (72.36%), followed by Logistic Regression (71.03%), Random Forest (70.74%), and KNN (69.73%). The results show that these models have the max precision for the given precision and recall.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 370-387 March 2026-Sept 2026 www.csjournals.com

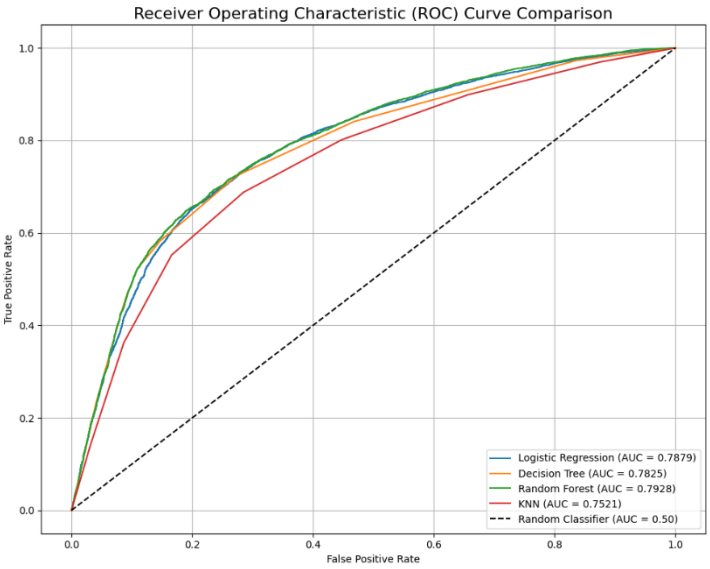

Figure 9: ROC Curves for All Four Models

The comparison of the models by the ROC curve is shown in figure 9. Random Forest got the highest AUC (0.7928), then Logistic Regression (0.7879), and after Decision Tree (0.7825), and KNN (0.7521). The performance of the overall classification of the models was compared in the following way: Random Forest was found to be better than the other models.

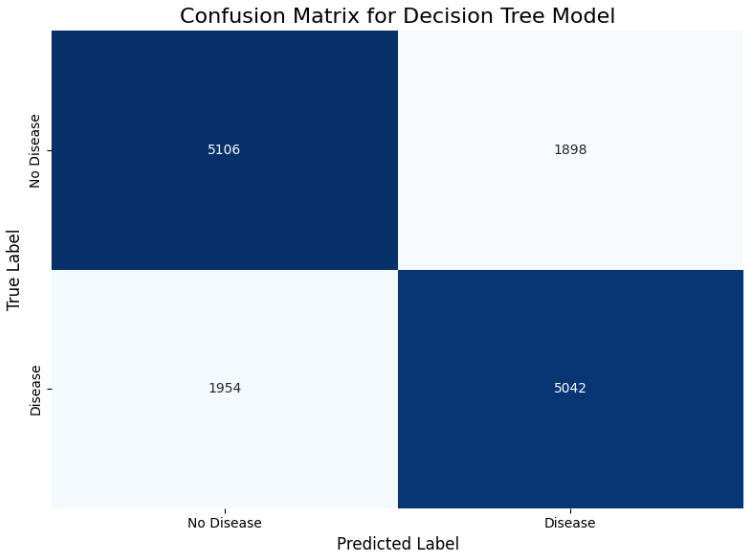

Figure 10: Confusion Matrix for Decision Tree

Figure 10 presents the confusion matrix of the Decision Tree model. The model got a correct classification rate of 5106/5106 (non-disease cases) and a correct classification rate of 5042/5042 (disease cases), with 1898/1898 (non-disease cases) and 1954/1954 (disease

cases) being misclassified. The overall results show good classification performance with balanced classification with good predictive capability.

6.1 Discussion

The results indicate that none of the classifiers was better than the rest with respect to all evaluation measures. Random Forest gave the best accuracy (72.69%), ROC-AUC (0.7928%) and precision (0.7612%) and worked well for the reliable classification task. Its recall (0.6607), however, was the lowest, implying that there were some missed positive cases.

By recall score (0.7207) and F1-score (0.7236), Decision Tree model is the best model for cardiovascular disease screening. Logistic Regression also gave a good performance with 72.49% accuracy and minimum train test gap suggesting good generalisation. Mridha et al. [4] found the same results. In case of KNN, it was unfit due to overfitting, which is in accordance to Teja and Rayalu [6].

Feature importance analysis showed that weight, height, systolic blood pressure and age were the most important features – consistent with known cardiovascular risk factors. Despite the outcomes of less accuracy compared with the papers presented by Santhosh et al. [1] and Sianga et al. [2] and Kumar and Lakshmanan [5].

7. Conclusion

The project used an easily replicable machine learning system that was developed from a patient record database of 70,000 patients to predict the onset of heart diseases early in life.

For test accuracy (72.69%), and ROC-AUC (0.7928), Random Forest was best describing all models, and for recall (0.7207) and F1-score (0.7236), Decision Tree was found best explaining all models, and hence was best suited for the screening of the disease. Logistic Regression captured the feature levels and was the better-performing model that was easily understandable; whilst K-Nearest Neighbours (KNN) was the least successful model due to overfitting.

Demographic characteristics whose weight was known to predict cardiovascular disease were included such as age, weight, blood pressure (BP) systolic BP and height. Finally, early prediction of cardiovascular diseases by the proposed strategy is reliable, explainable and it may play a role of decision making at early stages of the disease.

References

[1] S. Santhosh, K. Chadaga, R. V. Arjunan, and S. D'Souza, "Cardiac Clarity: Harnessing Machine Learning for Accurate Heart-Disease Prediction," IEEE Access, vol. 13, pp. 97529–97544, May 2025. DOI: 10.1109/ACCESS.2025.3573760

[2] B. E. Sianga, M. C. Mbago, and A. S. Msengwa, "Predicting the Prevalence of Cardiovascular Diseases Using Machine Learning Algorithms," Intelligence-Based Medicine, vol. 11, p. 100199, Jan. 2025. DOI: 10.1016/j.ibmed.2025.100199

[3] M. Almutairi and S. Dardouri, "Intelligent Hybrid Modelling for Heart Disease Prediction," Information, vol. 16, no. 10, p. 869, Oct. 2025. DOI: 10.3390/info16100869

[4] K. Mridha, M. Shukla, B. Acharya, V. C. Gerogiannis, A. Kanavos, M. A. Priyok, and R. H. Razu, "Implementing a Heart Disease Prediction Model with Explainable Machine Learning Techniques," SN Computer Science, vol. 6, p. 861, Sept. 2025. DOI: 10.1007/s42979-025-04409-z

[5] H. K. M. and S. Lakshmanan, "Comparative Analysis of Classification Models for Heart Disease Prediction Using Machine Learning," in Proc. International Conference on Inventive Computation Technologies (ICICT-2025), IEEE, 2025. DOI: 10.1109/ICICT64420.2025.11005114

[6] M. D. Teja and G. M. Rayalu, "Optimising Heart Disease Diagnosis with Advanced Machine Learning Models: A Comparison of Predictive Performance," BMC Cardiovascular Disorders, vol. 25, p. 212, Mar. 2025. DOI: 10.1186/s12872-025-04627-6

[7] H. Albanna, M. R. T. Tamang, C. Patel, and M. S. Sharif, "Revolutionising Heart Attack Prognosis: Introducing an Innovative Regression Model for Prediction," Informatics in Medicine Unlocked, vol. 57, p. 101664, Jun. 2025. DOI: 10.1016/j.imu.2025.101664

[8]Cardiovascular Disease Dataset, Kaggle. [Online]. Available: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset. [Accessed: 2024].

[a)] Corresponding Author: poojaa.sharma126@gmail.com